Introduction to R for Epidemiologists

Jenna Krall, PhD

Thursday, Feburary 12, 2015

Class survey

- Class is challenging
- Labs are useful (but long)
- You learned a lot from the homework, but it was difficult and took time
- Lecture portion needs to be revised
- 55.6% response rate

Sources of help

1. Lecture notes and labs
2. R help files
3. Short R reference card
4. Google
5. Instructor/TAs

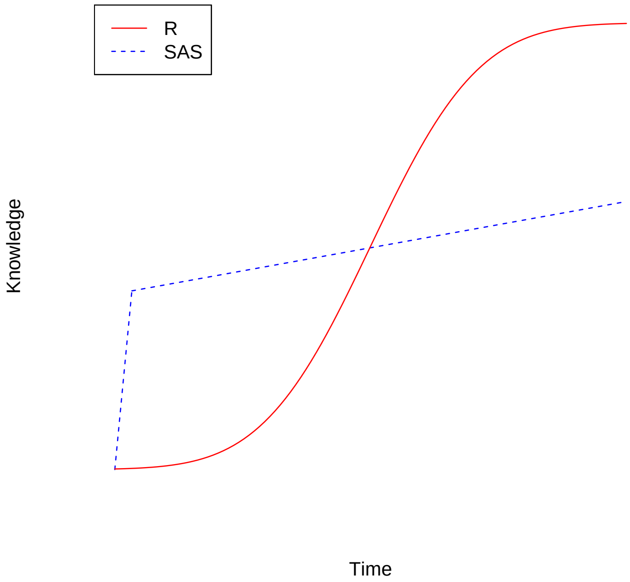
Class survey

Changes:

1. New lecture format
2. Shorter labs (with hints)
3. Shorter homeworks

Why are we here?

R is hard at the beginning



Outline

1. One sample T-tests
2. Two sample T-tests
3. Tests of proportion
4. Chi-squared tests
5. Relative risk
6. Odds ratio

One sample T tests in R

Review

- One sample Z and T tests are used for determining whether the mean in a population is different than a hypothesized value
- Examples
  - Is the average concentration of particulate matter air pollution in Atlanta different than 12 μ g/m^3?
  - Is the average gestational age for infants born with very low birthweight less than 39 weeks?

Assumptions for Z and T tests

- Large sample size or data are approximately normal if sample size is small

Assumptions for Z test

- Population standard deviation is known

One sample T-tests in R

Is average gestational age in the population different than 39 weeks (use α=0.05)?

- Null hypothesis $H_0: \ \mu = 39$
- Alternative hypothesis $H_1: \ \mu \neq 39$

One sample T-tests in R

```r
t_age <- t.test(x = vlbw$gest, mu = 39)
t_age
```

```
##
##  One Sample t-test
##
## data:  vlbw$gest
## t = -54.27, df = 173, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 39
## 95 percent confidence interval:
##  28.93 29.64
## sample estimates:
## mean of x
##     29.28
```

We reject the null hypothesis that the average gestational age of infants born with very low birthweight is significantly different than 39 weeks at α=0.05.

Two sample T-tests

Unpaired two sample t-tests

Recall that a two sample t-test tests the hypothesis that the means in two populations are the same:

- Is the average concentration of particulate matter air pollution in Atlanta different than the average air pollution concentration in Birmingham?
- Does the average gestational age of infants born with very low birthweight differ between males and females?

So we are testing whether the means of a continuous variable differ between two groups:

- Null hypothesis $H_0 : \mu_1 = \mu_2$
- Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Two sample T-tests

Paired two sample t-tests

- If the data are paired, use paired tests
  - e.g. Is the mean BMI the same after enrollment in an exercise program?
  - Paired tests account for the fact that we expect pairs to be more similar than we would expect if the data were unpaired.

Two sample T-tests

Does mean gestational age differ between male and female low birthweight infants?

```r
age_female <- vlbw$gest[vlbw$sex == "female"]
age_male <- vlbw$gest[vlbw$sex == "male"]
t.test(age_female, age_male, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  age_female and age_male
## t = -0.2063, df = 170.3, p-value = 0.4184
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##    -Inf 0.5191
## sample estimates:
## mean of x mean of y
##     29.24     29.32
```

Tests of proportion

We can also test proportions in R.

Is the proportion of those with pneumothorax different than 6.3%?

- One sample test of proportion
  - Null hypothesis $H_0: \; p_1 = 0.063$
  - Alternative hypothesis $H_1: \; p_1 \neq 0.063$

Is the proportion of those with pneumothorax different between multiple and singleton births?

- Two sample test of proportion
  - Null hypothesis $H_0: \; p_1 = p_2$
  - Alternative hypothesis $H_1: \; p_1 \neq p_2$

Tests of proportion

Is the proportion of pneumothorax different than 6.3%?

```
table_pneumo <- table(vlbw$pneumo)
table_pneumo
```

```
##
##   0   1
## 151  23
```

Tests of proportion

Is the proportion of pneumothorax different than 6.3%?

```
prop.test(matrix(c(127, 518), ncol = 2), p = 0.063)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  matrix(c(127, 518), ncol = 2), null probability 0.063
## X-squared = 193.6, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.063
## 95 percent confidence interval:
##  0.1673 0.2302
## sample estimates:
##      p
## 0.1969
```

Tests of proportion

Is the proportion of pneumothorax different between multiple and singleton births?

```
table_pneumo <- table(twin = vlbw$twn, pneumo = vlbw$pneumo)
table_pneumo
```

```
##      pneumo
## twin   0   1
##    0 115  17
##    1  36   6
```

```
table_pneumo <- matrix(c(95, 32, 415, 102), ncol = 2)
colnames(table_pneumo) <- c("Pneumo", "No pneumo")
rownames(table_pneumo) <- c("Not twin", "Twin")
```

Tests of proportion

Is the proportion of pneumothorax different between multiple and singleton births?

```
prop.test(table_pneumo)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table_pneumo
## X-squared = 1.533, df = 1, p-value = 0.2157
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.13695  0.03189
## sample estimates:
## prop 1 prop 2
## 0.1863 0.2388
```

Chi-squared test

Are two cateogorical variables independent?

- Is HIV infection associated with MRSA infection?
- Is sex associated with being a twin in very low birthweight infants?

Hypothesis:

- Null hypothesis: Sex is independent of being a twin
- Alternative hypothesis: Sex is not independent of being a twin

Assumptions:

- If 2x2 table, no cell counts < 5
- If rxc table, no more than 20% cells < 5

Chi-squared test

```r
chsq_surgery <- chisq.test(vlbw$sex, vlbw$twn)
chsq_surgery
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vlbw$sex and vlbw$twn
## X-squared = 0.3807, df = 1, p-value = 0.5372
```

```r
names(chsq_surgery)
```

```
## [1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
## [7] "expected"  "residuals" "stdres"
```

```r
chsq_surgery$p.value
```

```
## [1] 0.5372
```

Relative risk and odds ratio

Relative risk (RR)

- Ratio of risks: p_1/p_2
- Is the risk of disease the same in the exposed and unexposed groups?
- Often interested in testing $H_0: \; RR = 1$ vs. $H_1: \; RR \neq 1$
- Can only be calculated in prospective studies

Relative risk and odds ratio

Odds ratio (OR)

- Ratio of odds
- Is the odds of disease the same in the exposed and unexposed groups?
- Odds is **NOT** the same as risk
- Odds: p/(1-p) or p/q
- OR = $(p_1/(1 - p_1))$ / $(p_2/(1 - p_2))$ = (p_1/q_1) / $(p_2$ / $q_2)$
- Often interested in testing $H_0: \ OR = 1$ vs. $H_1: \ OR \neq 1$
- Useful in retrospective studies

Relative risk and odds ratio

```
library(epitools)
epitab(table_pneumo, method = "oddsratio")
```

```
## $tab
##          Pneumo    p0 No pneumo     p1 oddsratio  lower upper p.value
## Not twin     95 0.748       415 0.8027    1.0000     NA    NA      NA
## Twin         32 0.252       102 0.1973    0.7297 0.4627 1.151  0.1807
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

Relative risk and odds ratio

```
epi_pneumo <- epitab(table_pneumo, method = "riskratio")
epi_pneumo
```

```
## $tab
##          Pneumo     p0 No pneumo     p1 riskratio  lower upper p.value
## Not twin     95 0.1863       415 0.8137    1.0000     NA    NA      NA
## Twin         32 0.2388       102 0.7612    0.9354 0.8434 1.037  0.1807
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

Relative risk and odds ratio

```r
names(epi_pneumo)
```

```
## [1] "tab"         "measure"     "conf.level" "pvalue"
```

```r
epi_pneumo_out <- epi_pneumo$tab
colnames(epi_pneumo_out)
```

```
## [1] "Pneumo"     "p0"          "No pneumo" "p1"          "riskratio" "lower"
## [7] "upper"      "p.value"
```

Sample size calculations in R

How many observations would we need to test whether two means are different if

- The difference in means is 0.1
- The standard deviation is 1
- We want 90% power

```r
power.t.test(delta = 0.1, power = 0.9, type = "two.sample",
  alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 2102
##           delta = 0.1
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Survival data

Hunger games survival analysis: Do "career" tributes survive longer?

"which covariates are associated with the odds (or hazard ratios) being ever in your favor?"

- `http://www.bdkeller.com/writing/hunger-games-survival-analysis/`

(*source: Brett Keller*)

Survival data

```
library(survival)
hunger <- read.csv("Hunger Games survival analysis data set - Sheet1.csv",
  stringsAsFactors = F)
surv_hunger <- Surv(time = hunger$survival_days,event = rep(1, nrow(hunger)))
plot(survfit(surv_hunger ~ hunger$career),
  main = "74th annual Hunger Games - survival estimates",
  xlab = "Days", ylab = "Proportion surviving", col = c(1, 2))
legend(c("topright"), legend = c("Career tribute", "Not career tribute"),
    col= c(1, 2), lty = 1)
```

Survival data